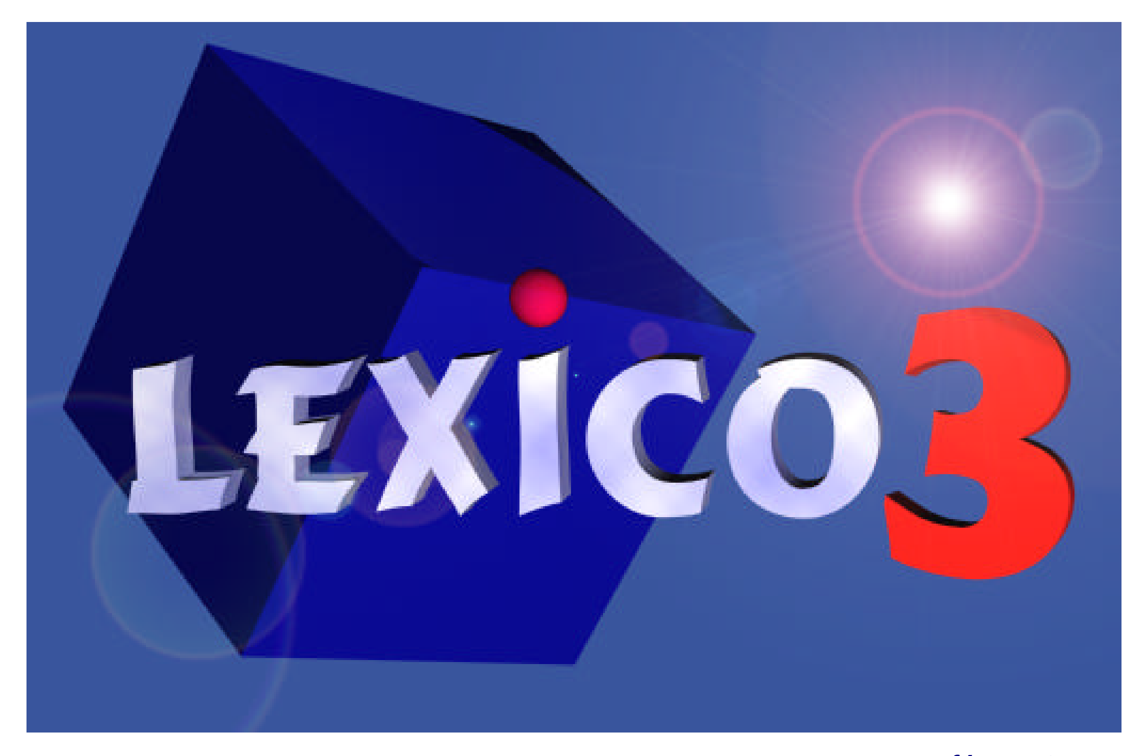
SYLED - CLA2T

Université de la Sorbonne nouvelle - Paris 3



Version 3.41 février 03

Outils de statistique textuelle

Cédric Lamalle
William Martinez
Serge Fleury
André Salem

Manuel d'utilisation

Béatrice Fracchiolla
Andrea Kuncova
Aude Maisondieu

TABLE DES MATIERES

Avant-propos

Lexico3 est l'édition 2001 du logiciel ***Lexico*** dont la première version remonte à 1990. Les fonctionnalités présentes dès la première version (segmentation, concordances, décomptes portant sur les formes graphiques, spécificités et analyses factorielles portant sur les formes et les segments répétés) ont été conservées et, la plupart du temps notablement améliorées.

L'originalité principale de la série ***Lexico*** est qu'elle permet à l'utilisateur de garder la maîtrise sur l'ensemble des processus lexicométriques depuis la segmentation initiale jusqu'à l'édition des résultats finaux. Les unités qui seront ensuite automatiquement décomptées sont exclusivement constituées à partir de la liste des délimiteurs fournie par l'utilisateur, sans recours à des ressources dictionnairiques extérieures.

Au-delà du repérage des seules formes graphiques, le logiciel permet d'étudier dans les textes la répartition d'unités plus complexes composées de séquences de forme : *segments répétés, couples de forme en cooccurrence, etc.* au contenu souvent moins ambigu que les formes graphiques dont elles sont composées.

Principales nouveautés

Une version orientée "objet"

La principale amélioration apportée à cette nouvelle version concerne l'architecture "objet" du programme. Les différents modules qui communiquent ensemble sont désormais capables d'échanger des données plus complexes (formes, segments répétés, cooccurrences dans l'avenir)..

Ainsi, il est désormais possible d'envoyer vers le module ***concordance***, comme vers tous les autres modules, des unités constituées dans les modules de ***segments répétés***, des listes de formes et de segments constituées dans les modules de ***spécificités***, etc. Ces possibilités permettent d'envisager une véritable *navigation lexicométrique*.

Constitution de groupes de formes

L'étude des accidents qui surviennent dans la répartition d'une forme graphique pour les différentes parties d'un corpus de textes, suscite inévitablement des questions à propos de la répartition d'autres unités graphiques qui lui sont liées au plan linguistique (autres réalisations du même lemme, formes liées au plan sémantique). De nouveaux outils (recherche des expressions régulières) ont été intégrés qui facilitent la recherche de tels ensembles de formes.

Localisation des particularités lexicométriques

La caractérisation des différentes parties d'un corpus par les formes qu'elles emploient abondamment est rendue plus précise dans la présente version par la possibilité de mettre en évidence des sections du texte dans lesquelles telle particularité de répartition est particulièrement remarquable. La matérialisation de ces sections sur des diagrammes représentant le texte permet de dresser une véritable topographie textuelle.

Pour en savoir plus

En ce qui concerne les modifications, les correction des erreurs, les mises à jour, la source principale est le site ***Lexico3*** de l'équipe SYLED-CLA2T à l'université de la Sorbonne nouvelle-Paris3.

http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

On trouve sur ce site les versions antérieures de ***Lexico*** (***Lexico1*** - MacIntosh, ***Lexico2*** PC) ainsi que divers documents téléchargeables, parmi lesquels le présent manuel.

Une bibliographie générale est donnée en annexe. Les renvois à l'ouvrage :
Lebart Ludovic, Salem André, ***Statistique textuelle,*** Dunod, Paris 1994,
sont notés (*L&S,* p. xxx)

Développements à venir

Certaines procédures couramment utilisées dans les recherches lexicométriques n'ont pu être intégrées dans la présente version.
C'est le cas, par exemple, pour la Classification Ascendante Hiérarchique (CAH) ainsi que pour certaines méthodes permettant de mettre en évidence les réseaux de cooccurrences dans un texte. Ces procédures seront disponibles dans la prochaine version de Lexico.

Installation

0.1 Avertissement

Il se peut, malgré tout le soin que nous avons apporté à la préparation de cette version, que quelques erreurs subsistent encore. Nous vous demandons de signaler les éventuelles anomalies à :

Lexico3 / ILPGA : 19, rue des Bernardins 75005 Paris, France

Joindre à l'envoi le corpus de texte sur lequel des dysfonctionnements auront été constatés ainsi que le fichier *atrace.txt,* créé au moment de l'exploitation dans le répertoire où se trouve le corpus analysé, qui contient des renseignements indispensables pour le débogage.

Configuration minimale

A partir de Windows 95

Processeur 486, 4Mo de mémoire vive

3 Mo libres sur le disque dur

Configuration conseillée

Windows 98 et postérieurs

Pentium II, 8Mo de mémoire vive

3 Mo libres sur le disque dur

Lexico3 fonctionne sous Windows 95 et postérieurs, et Windows NT 3.51 et 4.0.

Nous conseillons vivement de regrouper programme et corpus dans un même dossier sur le disque dur.

0.2 Installer le logiciel

Pour installer LEXICO3
Insérer le CD ROM dans le lecteur.
Double-cliquer sur l'icône du fichier SETUP.EXE qui se trouve sur ce CD-ROM.
Suivre les indications données par le programme d'installation
Le message : ***Lexico3 a été installé*** signale la fin de l'installation.

1 Les corpus de texte

L'analyse lexicométrique compare les décomptes réalisés à partir du repérage des occurrences d'unités lexicales (formes, segments, types généralisés, etc.) dans les différentes parties d'un corpus de textes.

Cette introduction s’ouvre sur des exemple élémentaires (section 1.1) permettant d'aborder rapidement le logiciel. Les problèmes concernant la segmentation automatique sont présentés ensuite (section 1.2). La section (section 1.3) présente le cas d’un corpus en grandeur réelle.

Prise en main rapide

Les deux sections qui suivent s'adressent aux utilisateurs désireux d'explorer rapidement les principales fonctionnalités du logiciel.

Corpus d'initiation auteurs.txt

A partir du fichier d’initiation ***auteurs.txt***, qui se trouve sur le CD, on peut réaliser une partition en trois parties qui permettra ensuite d’effectuer des comparaisons entre les « textes » rassemblés dans ce *corpus* .

Exemple de balisage d'un corpus **: le fichier *auteurs.txt***

```
<Auteur=Nerval>
Il est un air pour qui je donnerais tout Rossini tout Mozart tout Weber
<Auteur=Trenet>
Y a d' la joie ! Bonjour bonjour, les hirondelles ! Y a d' la joie, dans le
ciel par-dessus le toit
<Auteur=Brassens>
```

La canne de Jeanne est morte au gui l'an neuf, elle avait fait la veille merveille un oeuf

La clé ***Auteur*** permet ici de partager le corpus en trois parties qui seront ensuite comparées entre elles

Pour réaliser cette opération, il faut :

?? Ouvrir ***Lexico3***, en cliquant sur l’icône du logiciel**.**

?? Sélectionner le fichier à ouvrir dans le menu **Fichier (**ici **auteurs.txt)**

?? Accepter les paramètres de segmentation qui seront définis plus bas en cliquant sur le bouton OK.

Lexico3 présente alors dans la partie gauche de l’écran, la liste des formes attestées dans le corpus munies de leurs fréquences respectives. On peut alors effectuer toute une série d’opérations lexicométriques décrites dans la suite de ce manuel en actionnant les boutons qui appellent les différents modules du logiciel (cf. sections 2-4).

Votre corpus d'essai

Comme dans l'exemple précédent, insérer quelques *balises* pour délimiter différentes parties dans le corpus (par exemple : <partie=1>, <partie=2>, etc..

Dans le dossier Lexico3 créé lors de l'installation du logiciel, à l'aide de votre traitement de texte habituel (Word, etc.), enregistrer votre document avec l'option ***texte seulement*** (article ***Enregister sous*** ... du menu ***Fichier***).

Votre corpus d'essai est prêt pour l'exploitation par ***Lexico3***. Pour une première utilisation, le plus simple est d'accepter les paramètres de segmentation *par défaut* proposés par le logiciel (caractères délimiteurs etc.).

1.2 Normes d'enregistrement

Dans la période récente, de nouvelles normes (XML, HTML, etc.) tendent à s'imposer pour le stockage des corpus textuels. Cependant, les corpus réunis pour l'analyse lexicométrique sont encore constitués de documents provenant de sources différentes, souvent stockés sous des formats variables. Pour éviter de mettre en évidence des variations entre les textes qui renvoient à des méthodes de stockages différentes, il est utile de soumettre les textes à un travail de normalisation minimal. Plusieurs logiciels (dont ***MKCorpus***[1], fourni sur le CD-ROM), prennent en charge une partie du nécessaire travail d'homogénéisation.

L'analyse lexicométrique étudie la répartition dans les textes d'unités complexes (*lemmes*, *segments répétés*, *cooccurrences*, *types généralisés*). Cependant, une segmentation en formes graphiques constitue une première étape nécessaire qui permet tout à la fois :

?? d'obtenir une première estimation des principales caractéristiques lexicométriques du corpus (nombre d'occurrences, de formes, d'hapax, fréquence maximale);

?? de réaliser les premières typologies sur les parties du corpus;

?? de localiser les erreurs qui subsistent après les premières corrections.

Pour réaliser cette segmentation en formes graphiques, il faut définir des normes. Dans le cas de ***Lexico3***, ces normes sont particulièrement simples.

Le texte doit être enregistré sous la forme d'un fichier ***texte seulement*** (*.txt)[2].

Délimiteurs / non-délimiteurs

Dans le corpus soumis aux traitements lexicométriques, une forme graphique est une suite de caractères non-délimiteurs, encadrée par deux caractères délimiteurs. Cela veut dire que les formes graphiques -dont on va décompter les occurrences- sont entièrement définies par la liste des délimiteurs retenus par l'utilisateur. L'*identification* se produit lorsque les chaînes

---

[1] ***MKCorpus*** est développé par S. Fleury (Paris3 -Ilpga - Syled).

[2] On écarte les fichiers de type document (*.doc) et autres formats créés par traitement de texte, car ceux-ci intègrent un en-tête renfermant diverses informations, sur la mise en forme notamment.

comprises entre deux délimiteurs de formes sont identiques. Ainsi, si l'on n'effectue pas de prétraitement approprié, *Vache* n'est pas identique à *vache*, et *c'est-à-dire* est différent de *c'est à dire*.

La partie mécanique de la segmentation automatique est considérablement simplifiée par la réalisation du principe simple :

un signe = un statut

Chaque signe typographique doit être susceptible de recevoir un statut (*délimiteur* ou *non délimiteur*) fixé une fois pour toutes au début de la procédure.

Ces principes entrent parfois en conflit avec les conventions typographiques usuelles. Le tiret de *dit-il* n'a pas le même statut grammatical que celui de *garde-manger*. Il en va de même pour l'apostrophe de *aujourd'hui* (qui devrait être considéré comme *non-délimiteur*), dont le statut diffère de celle qu'on trouve dans la séquence *j'aime*.

Lexico3 propose une liste de délimiteurs par défaut qui peut être modifiée par l'utilisateur: `-—_:;/.,?¿!¡*$"+=(){}.` L'espace (blanc) est automatiquement ajouté à cette liste. Une fois la liste des délimiteurs fixée, les autres caractères : `a, b, c,...` deviennent des *caractères non-délimiteurs*.

Toute suite de caractères non-délimiteurs, bornée à ses deux extrémités par des délimiteurs, est considérée comme l'occurrence d'une forme à relever et à classer.

Minuscules, majuscules, apostrophes

Pour des visées particulières, l'utilisateur peut combiner les conventions de pré-traitement et les options de segmentation pour influer sur le type des formes produites par la segmentation. On peut par exemple remplacer systématiquement, lors d'un traitement préalable, toutes les majuscules du texte par une astérisque suivie de la minuscule correspondante (ex : *Moi* devient **moi*). Une segmentation intégrant le caractère * parmi les délimiteurs confondra les occurrences des séquences *Moi* et *moi*; une segmentation pour laquelle l'astérisque n'appartient pas à cette liste produira des décomptes distincts pour les deux séquences.

Sections du texte

Au-delà des partitions logiques repérables dans le texte, celui-ci contient des marques de rupture qui constituent sa respiration (phrases, paragraphes, etc.). ***Lexico3*** offre la possibilité de promouvoir un caractère délimiteur (ou plusieurs d'entre eux) au rang *de délimiteur de section*. Ce précodage permet d'étudier ensuite la répartition des occurrences d'une unité lexicométrique parmi les sections ainsi constituées.

NB : L'insertion systématique de caractères délimitant des sections peut être réalisée en utilisant la fonction ***Remplacer*** d'un traitement de texte[3].

---

[3] On changera ici systématiquement les caractères retour-chariot par la séquence retour-chariot+blanc+caractère §.

Clés /Balises

Au cours d’une étude lexicométrique, on cherche à comparer les fréquences des formes dans les différentes parties d'un corpus. Pour rendre possible ces comparaisons, le texte doit comporter des balises indiquant les délimitations logiques du corpus.

Les parties définies par l’utilisateur peuvent être chronologiques, comme dans l’exemple du *Père Duchesne*, (cf. section 1.2, « Prise en main rapide ») mais aussi thématiques.

Coder une clé.
Une clé (*ex :* <Auteur=Dupond>) se compose de 5 éléments :

1	<	un chevron ouvrant
2	**Auteur**	le **type** de la clé
3	=	le signe "égal"
4	**Dupond**	le **contenu** de la clé
5	>	un chevron fermant

Exemples : <Année=1998>, <Auteur=Jean_de_la_Fontaine>

L’insertion de *clés* constitue une phase importante dans la préparation du texte. Les clés introduites permettront ensuite à l’utilisateur d'effectuer des comparaisons à partir des parties du corpus qu'elles découpent.

1.3 Normes de dépouillement

Pour procéder à des dépouillements statistiques à partir des textes ainsi stockés, il faut définir des normes d'identification des unités textuelles. Comment identifier les occurrences d'un même type au fil du texte ? Plusieurs normes sont envisageables qui s'appuient chacune sur des savoirs, des pratiques, des perspectives différentes.

- ✍ Le dépouillement en ***formes graphiques*** (identification automatique des occurrences d'une même chaînes de caractères) est particulièrement facile à décrire et à mettre en œuvre.
- ✍ Le dépouillement en ***lemmes***, s'appuie sur des ressources extérieures (dictionnaires de lemmes, analyseurs syntaxiques).
- ✍ Certains logiciels proposent également le regroupement d'occurrences qui peuvent être rapportées à une même ***racine*** ou ***n-gramme*** à l'aide de processus plus au moins automatisés.

Au-delà du simple dépouillement en formes graphiques, ***Lexico3*** permet de recenser différents types d'unités textuelles.

- ✍ Les ***segments répétés*** : suites de formes graphiques identiques attestées plusieurs fois dans le texte).

- ✍ Les ***cooccurrences*** : couples de formes présentes dans les mêmes contextes (phrase, sections, etc..
- ✍ Les ***types généralisés*** ou ***Tgen(s)***: unités de dépouillement définies par l'utilisateur à l'aide d'outils lui permettant d'effectuer automatiquement des regroupements d'occurrences du texte (ex : les occurrences des formes qui commencent par la séquence de caractère *patr : patrie, patriotes, patriotisme, etc.*).

1.4 Exemple : le corpus *Duchesne*

Text1.txt est un fichier contenant un fragment du corpus *Père Duchesne*[4] (Duchn.txt). Les deux fichiers sont disponibles sur le CD-Rom d'installation du programme.

Tableau 1.1 : Exemple de codage de corpus

```
<An=1793> <Numero=220> <S03=0> <Epg=1>
```

[4] Le corpus *Père Duchesne* réuni par Jacques Guilhaumou dans le cadre du laboratoire *Lexicométrie et textes politiques* de l'ENS de Fontenay/St. Cloud a fait l'objet de nombreuses études, notamment des études de caractère méthodologiques (cf. blibliographie infra).

```
§ la grande colère du *père *duchesne , de voir que les
mouchards de *la-*fayette et tous les fripons soudoyés par la
liste civile, veulent rétablir les compagnies de grenadiers et
de chasseurs, pour égorger les *sans-culottes et les chasser
des assemblées de *section .ses bons avis aux *lurons des
*faubourgs pour qu' ils arrachent les moustaches postiches à
ces grenadiers de la vierge *marie , qui veulent rétablir la
royauté.
<S03=1>
§ millions de tonnerre, nous ne mettrons donc jamais les
fripons à la raison ? ils <Epg=2>ont laissé tomber leurs
masques et nous les voyons à nu. serons nous encore dupes des
fripons? quand je voulais faire la conduite de *grenoble à
tous les talons rouges quand je disais, du soir au matin, que
tous les ci-devant ne cesseraient de nous trahir, n' avais je
pas raison, foutre?
§ je me suis toujours plus défié des nobles convertis que des
émigrés. c' est pour nous frapper de plus près que ces gredins
sont restés au milieu de nous. ils ont fait les chiens
couchants pour mieux nous tromper. jamais, foutre, ils n' ont
cessé de s' entendre avec les ennemis du dehors. ce sont eux
qui nous ont mis à chien et à chat, qui ont brouillé les
cartes dans les trois assemblées nationales, et corrompu les
représentants du peuple. si nous avions eu assez d' estoc pour
les envoyer tous à *coblentz au commencement de la révolution,
nous n' aurions pas acheté notre liberté par des flots de
sang; nous aurions depuis longtemps une constitution; la paix
et le bonheur régneraient dans notre république.
```

Dans ces fichiers-textes fournis à titre d'exemple,

- la clé Sda permet de coder l'année durant laquelle le texte a été publié.
- la clé Numero permet d'introduire un numéro de livraison qui respecte l'édition originale du texte (96 livraisons numérotées de 255 à 351 pour le corpus DUCHn.txt, 6 numéros pour le sous-corpus text1.txt).
- la clé Epg permet le passage à une autre page, conformément à la pagination de l'édition originale du corpus.
- la clé S03 permet de distinguer les portions de texte qui sont des titres et des chapeaux (S03=0) du texte proprement dit (S03=1).
- le caractère paragraphe § marque le début de chacun des paragraphes du texte.
- le caractère * permet d'identifier les majuscules du document original.

2 Outils d'exploration textuelle

Cette section décrit les fonctionnalités de ***Lexico3*** qui permettent de retrouver au fil des textes les occurrences des différentes unités textuelles que l'on peut construire à partir de la séquence textuelle (*formes graphiques*, *segments répétés*, *groupes de formes*, *Tgens*).

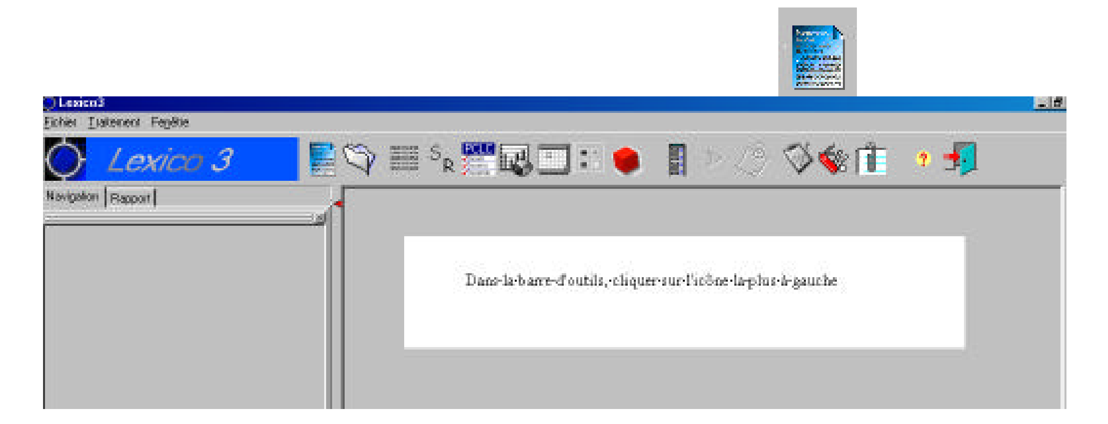
2.1 Segmenter un corpus

La segmentation crée une base de données textuelles à partir d'un corpus ***Moncorpus.txt*** fourni par l'utilisateur. Cette base se compose de trois fichiers (***Moncorpus.dic***, ***Moncorpus.par***, ***Moncorpus.num***) dont les deux premiers peuvent être visualisés au moyen d'un traitement de texte quelconque.

Mise en œuvre pratique

Ouvrir le logiciel en cliquant deux fois sur l'icône INSERER ICÔNE LEXICO3
Dans la barre d'outils, cliquer sur l'icône la plus à gauche





Cliquer sur cette icône pour ouvrir un fichier texte

Le programme propose de choisir un fichier texte dans un répertoire selon les procédures habituelles du système d'exploitation Windows.

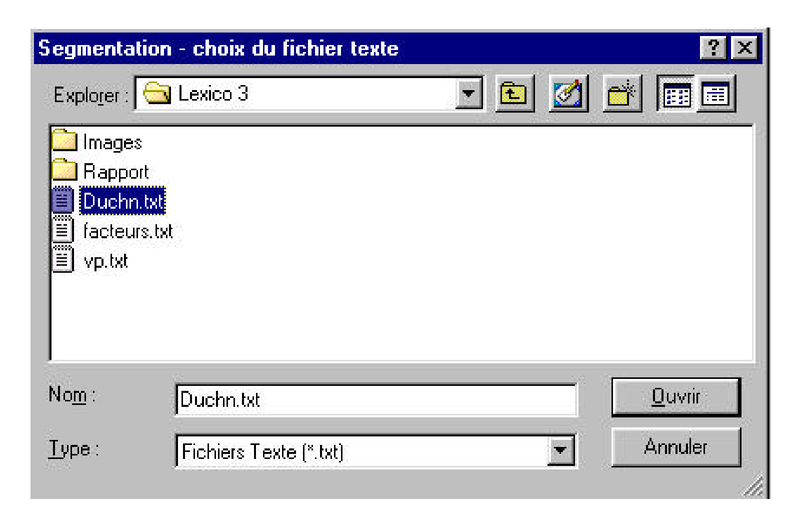


Figure 2.1 : Sélection d'un fichier texte

Sélectionner le fichier qui contient le corpus à segmenter *Duchn.txt*. Une boîte de dialogue apparaît qui permet de régler les paramètres de la segmentation à l'aide des délimiteurs (cf. 1-Préparation du texte).



Figure 2.2 : Boîte de sélection des paramètres de segmentation

Rappel : Il est possible de modifier cette liste des délimiteurs.
Lancer ensuite la segmentation en cliquant sur le bouton OK.

Vérification des clés

Le programme vérifie la conformité du corpus d'entrée aux normes décrites plus haut. Ce module signale notamment les clés mal codées :

clé non fermée	<S01=Aglaé
espace dans le type ou le contenu de clé	<S 01= Ag laé>

balise de fermeture isolée	elle est < belle.
absence de signe =	<S01Aglaé>
absence de contenu de clé	<S01=>
absence de type de clé	<=Aglaé>

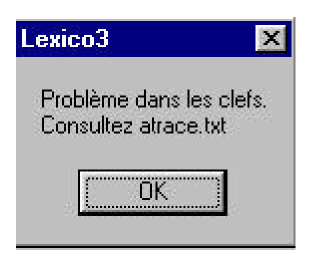


Figure 2.3 : Avertissement d'erreur de codage de clé

Des informations plus détaillées sur les erreurs sont fournies par le fichier de compte-rendu **`atrace.txt`**, dans le même dossier que le fichier texte, qui indique également le numéro de la ligne incriminée. Les erreurs apparaissent alors comme ci-dessous :

Tableau 2.1 : Compte-rendu de la segmentation
(Lxxx…. indique la ligne fautive)

```
*****COMPTE-RENDU DE LA SEGMENTATION*****
Fichier -- C:\LEXICO3T\TEXTES\DUCH.TXT -- ouvert pour vérification
L    2  Clé incorrecte :(espace dans contenu de clé) : <Sda=17 93>
L   94  Clé incorrecte :(pas de contenu de clé) : <Epg=>
L 5709  Clé incorrecte : Mauvais emplacement de balise de fermeture
L 5845  Clé incorrecte :(espace dans le type de la clé) : <Ep g=3>
L13277  Clé incorrecte :(mauvaise fermeture de la clé) <S02=330 <
L13496  Clé incorrecte :(pas de signe "=") : <Epg8>
```

Segmentation du texte

Lorsque les lignes fautives ont été corrigées, on relance le programme comme indiqué plus haut. S'il n'y a plus d'erreurs, une jauge permet de suivre la progression de la segmentation du texte.
A l'issue de la segmentation, la partie gauche de l'écran affiche la liste lexicométrique des formes du corpus, à côté de chacune de ces formes, on trouve sa fréquence dans l'ensemble du corpus. On appelle hapax toute forme n'ayant qu'une seule occurrence sur l'ensemble du corpus. On obtient un classement alphabétique de cette liste en cliquant sur le bandeau gris situé juste au-dessus du dictionnaire (ordre lexicographique). Un second clic remet la liste dans son état initial (ordre lexicométrique).

Fichiers de sortie

Plusieurs fichiers sont créés et enregistrés sur le disque dur dans le même répertoire que le texte-source. Si le corpus soumis à la segmentation s'appelle : *nomgénérique.txt*, ces fichiers s'appellent respectivement : *nomgénérique.par, nomgénérique.dic, nomgénérique.num.*
Le fichier *nomgénérique.par* : contient les principaux décomptes portant sur les formes, les occurrences, etc. ainsi que le rappel des caractères délimiteurs choisis lors de la segmentation.

Tableau 2.2 : Exemple de fichier paramètres (.par)

```
Lexico3.1 PC DUCH
nbetiq=0
196125 196125 11023 142185 10859 6130 4953 5000000 14 8 143 0 0
*** Résultat de la segmentation du fichier: DUCH.TXT ***
Délimiteurs #-—:;/\\.,?¿!¡*$\'"+=(){}[]§
nombre des occurrences : 142185
nombre des formes : 10859
frequence maximale : 6130
nombre des hapax : 4953
nombre des clés(type) : 8
nombre des clés(ctnu) : 143
*** Fin de la segmentation du fichier: DUCH.TXT ***
```

Le fichier *moncorpus.dic* : contient le dictionnaire des formes classées par fréquence (un enregistrement pour chaque forme). A côté de la fréquence de la forme on trouve son rang lexicographique dans le corpus (i.e. son numéro dans la liste des formes classées selon l'ordre lexicographique.
Le fichier *Moncorpus.num* : contient le texte numérisé, c'est-à-dire, sous une forme codée de façon compacte, les occurrences, des formes, ponctuations, clés et autres éléments du corpus. Ce fichier à usage interne ne peut être consulté à l'aide d'un éditeur de texte.
Le fichier ***atrace.txt*** contient un rapport détaillé des opérations effectuées par le programme (mémoire allouée, paramètres pris en compte, fichiers lus et écrits, ...). En cas d'échec du traitement, ce fichier peut fournir des indications permettant de cerner la cause des difficultés.

Tableau 2.3 : Extrait du dictionnaire

```
frq     rang lex.     forme
6130      2703         de
4749      6033         les
4298      5909         la
3773      4216         et
(…)       (…)          (…)
   1     10967         voyager
   1     10987         zeté
-----------------------------   Fin de la zone des formes graphiques
      259
    10859  !
198 10860  "
 49 10861  $
-----------------------------   Fin de la zone des ponctuations
766 10873 Epg
 96 10874 S01
-----------------------------   Fin de la zone des types de clés
97 10882 01
 1 10883 02
-----------------------------   Fin de la zone des contenus de clés
```

Tableau 2.4 : Extrait du fichier de trace (*atrace.txt*)

LecParam

192000 192000 11169 142177 10988 6130 5056 5000000 14 8 159
Allocation de la mémoire :
Allocation de lexm réussie, 178720 octets
Allocation de tnum réussie, 768000 octets
Allocation de ftext réussie, 446800 octets
Allocation de list réussie, 24520 octets
Entrée dans OpenDicNum
Dictionnaire numérisé : Duchn.dic
Entrée dans OpenTextNumFichier Texte : DUCH.num : 192083 items.
Fichier Param DUCH.par :

2.2 Ouverture d'une base déjà créée

On est souvent amené à faire des expériences sur un même corpus lors de sessions de travail espacées dans le temps. En réutilisant une *base* créée lors d'une session précédente on est sûr que l'on utilise bien, lors de la session ultérieure, les paramètres de segmentation fixés lors de la première session.

NB: Il est possible d'ouvrir un texte déjà segmenté en le glissant directement sur l'icône de ***Lexico3*.**

2.3 Concordances

L'outil ***Concordance*** permet de visualiser toutes les occurrences d'une *forme* ou d'un *type généralisé* (Tgen) en contexte. La concordance permet un retour systématique au texte et à l'environnement immédiat de la forme.

Sélectionner une forme (ou un type)

Cliquer sur l'icône ***Concordance***, une boîte de dialogue apparaît :
Pour obtenir la concordance d'une forme, on peut au choix :

- ? entrer la forme dans la zone d'édition 'forme pivot' (ex : homme), puis appuyer sur le touche 'Entrée'.
- ?? faire glisser cette forme sur la fenêtre des concordances à partir du dictionnaire ou à partir du *Garde-mots*,
- ?? faire glisser un lien réalisé à partir de la fenêtre *groupe de formes* ou un *segment répété* (voir section 2.5 'segments répétés') dont on souhaite étudier les contextes, puis la/le glisser dans la fenêtre de droite et lâcher. La concordance de toutes les occurrences du *Tgen* en contexte s'affiche alors automatiquement.
- ?? dans une fenêtre de concordance relative à une forme donnée, on peut sélectionner toute autre forme visible dans la fenêtre et obtenir sa concordance.

Lancer l'exécution de la requête en appuyant sur la touche *Entrée*. La liste de toutes les occurrences en contexte du *type* pour lequel l'on a lancé la recherche s'affiche à l'écran.

Glisser / Déposer
Sélectionnez une forme - clic gauche de la souris.
Maintenez le bouton gauche enfoncé et faites glisser la forme sélectionnée vers l'endroit souhaité puis déposez (lâchez le bouton gauche).

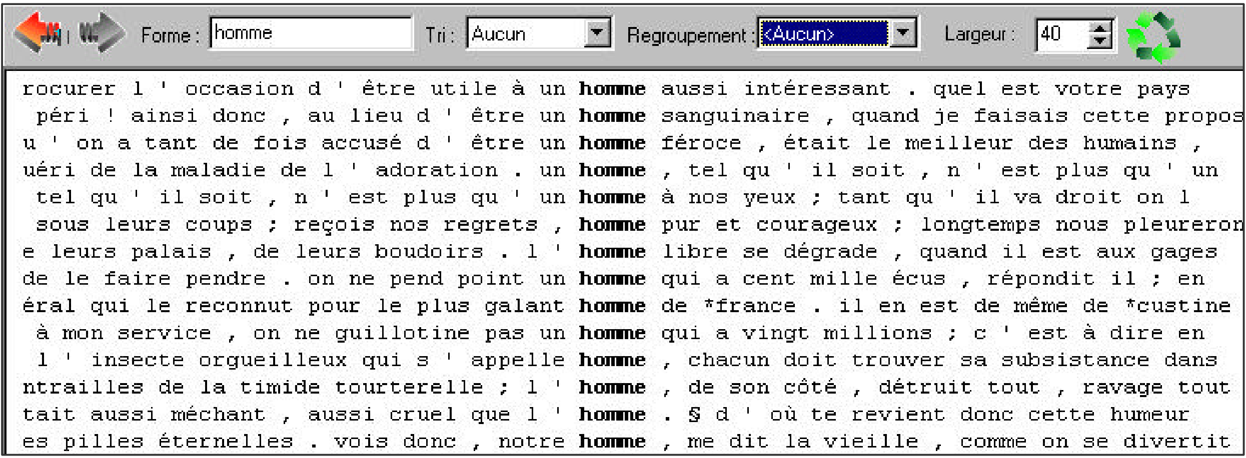


Figure 2.4: Concordances
Extrait d'une concordance autour de la forme-pôle *homme*
dans le corpus *Duchesne***.**

Possibilités d'affichage de la concordance

On choisit l'ordre de tri des contextes à partir du menu déroulant « tri » (avant, après, aucun).
La boîte déroulante ***Regroupement*** permet de regrouper les contextes en fonction d'une partition (par exemple, par locuteur, mois ou année).
?***Largeur*** : permet choisir le nombre de caractères (espaces inclus) qui doivent apparaître avant et après chaque pôle. Pour le modifier après une première recherche, changer la largeur et cliquer sur rafraîchir (**Figure 2.4**).

Les tris

Les différents contextes relatifs à une même forme peuvent être ordonnés de trois manières différentes. Le tri de ces contextes peut être effectué en fonction de :

- l'ordre alphabétique de l'occurrence qui précède la forme-pôle (tri avant)
- l'ordre alphabétique de l'occurrence qui suit la forme-pôle (tri après)
- l'ordre d'apparition des occurrences de la forme-pôle dans le texte.

Les bouton ***Précédent*** et ***Suivant*** (flèches rouges à gauche de la fenêtre), permettent de naviguer parmi les concordances réalisées pour différentes formes, types, etc.



2.4 Ajouter les résultats au rapport

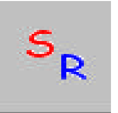
tous les documents produits par ***Lexico3***, chaque concordance peut être ajoutée au final.

Le rapport

Les résultats qui intéressent l'utilisateur pour une exploitation ultérieure peuvent être rassemblés dans un dossier nommé ***Rapport***. Ce dossier aisément manipulable à l'aide d'un navigateur web (*Internet Explorer*, *Netscape*, etc.) contient un fichier *index.htm* qui permet la navigation parmi les résultats sélectionnés. Le rapport peut être consulté à tout moment à la condition que l'utilisateur l'ait préalablement enregistré (cf. section 4.3).

Ajouter au rapport

Pour ajouter un document au rapport, il suffit de cliquer sur l'icône *Ajouter au rapport* décrite dans cette section. Dans le cas général, on utilise l'icône située dans la barre des outils. Pour certains documents (sections, listes, etc.), on utilise un bouton similaire situé dans la fenêtre correspondante.



2.5 Recherche des segments répétés

Les *segments répétés* sont des suites de formes dont la fréquence est supérieure à 2 dans le corpus[5]. On trouve par exemple dans le corpus ***Duchesne*** les segments

Segment	*longueur*	*fréquence*
tirer les marrons du feu	5	6

[5] (*L&S*, p. 58)

Pour créer la liste des segments répétés, cliquer sur l'icône ***SR***; une boîte de dialogue apparaît qui permet de paramétrer la sélection des segments répétés (**figure 2.5**):

La partie supérieure de la fenêtre permet de fixer le *statut* des *caractères délimiteurs* du texte (le statut par défaut est *délimiteur de séquence*. Pour modifier ce statut, annuler la coche en regard du caractère correspondant). Les segments répertoriés ne chevaucheront pas ce type de délimiteur.
La partie inférieure permet de décider du statut des ***clés*** rencontrées dans le corpus (Ici, par exemple, on permettra à un segment de chevaucher une clé indiquant un changement de page, mais non une clé indiquant un changement de partie).
On fixe une fréquence minimum en-dessous de laquelle les formes et les segments ne seront pas retenus. Ce seuil est fixé à 10, par défaut.
Le bouton OK permet de lancer la recherche des segments répétés.

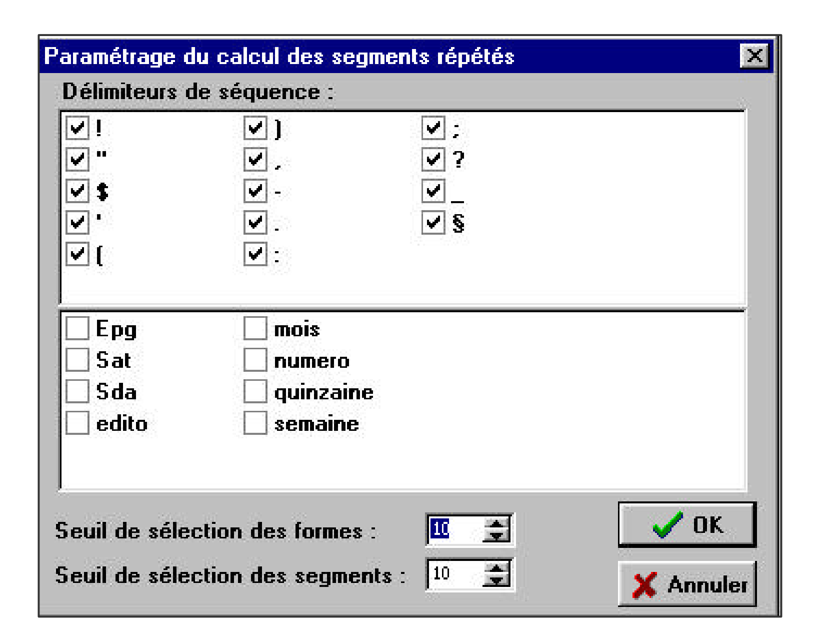


Figure 2.5 : Délimiteurs et seuil des formes

La liste des segments répétés répertoriés dans le texte apparaît dans la partie gauche de la fenêtre. Elle est consultable en cliquant sur l’onglet ***Segments répétés*** .

Segments répétés

Lg	Segment	Frq
2	de *brissot	14
2	de *coblentz	15
2	de *cobourg	10
2	de *custine	20
2	de *dumouriez	13
2	de *france	12
2	de *lyon	29
2	de *marat	13
2	de *marseille	23
2	de *paris	62
3	de *pitt et	10
2	de *pitt	17
2	de *toulon	10
2	de bataille	10
2	de bien	11
2	de bon	18
2	de bonne	14
2	de bons	18

Figure 2.6 : Liste des segments répétés

2.6 Groupe de formes

L'outil ***Groupe de formes*** permet de constituer des *types* rassemblant les occurrences de formes graphiques différentes liées par une propriété commune. On peut ainsi, moyennant certaines précaution, rassembler le pluriel et le singulier d'une même forme, les flexions d'un même verbe, des formes qui possèdent un lien sémantique, etc.. Les formes ainsi regroupées peuvent ensuite être manipulées comme des entités uniques les *Tgen*.

On lance simultanément une recherche sur plusieurs formes, en introduisant des chaînes de caractères qui correspondent à des préfixes, des suffixes ou des suites de caractères graphiques.

Mise en œuvre

– Entrer le nom du groupe de formes.
– Entrer la forme à rechercher.
– Cliquer sur rechercher.

L"objet" résultant peut ensuite être manipulé comme une forme "classique", en cliquant sur la flèche rouge du groupe (et en maintenant le clic gauche), on "glisse" le groupe sur la carte de la partition. (**Figure 2.7)**
Lors d'une nouvelle recherche, les nouveaux résultats se concatènent aux précédents.

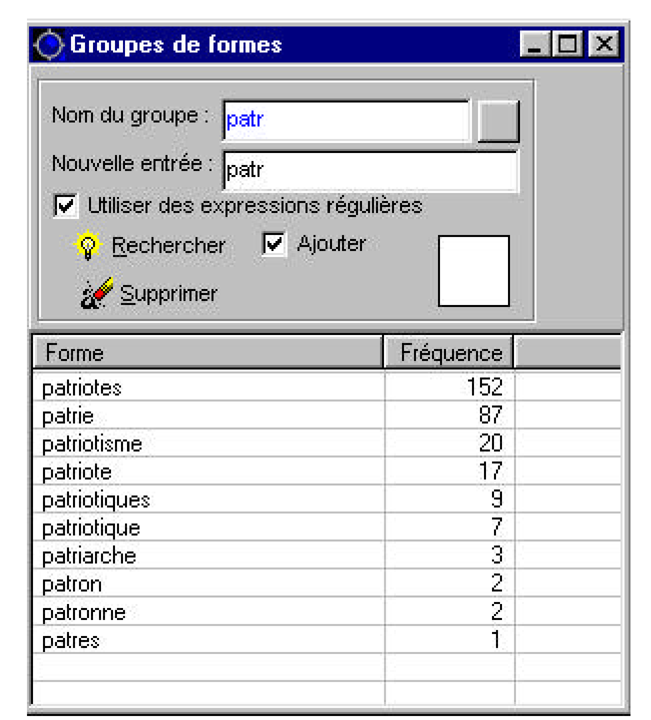


Figure 2.7 : Création de groupes de formes

Le bouton ***Supprimer*** permet d'affiner cette liste en éliminant, par exemple, après les avoir sélectionnées, les formes *patriarche*, *patron*, *patronne*, *patres*, etc.

Expressions régulières

Nous avons retenu un langage d'*expressions régulières* (ou *rationnelles*) couramment utilisé dans le monde de l'informatique pour permettre à l'utilisateur de constituer des groupes[6].

Pour rechercher des formes (Tgen) -via les expressions régulières- ***Lexico***, va effectuer, par défaut, une recherche de mot commençant par la chaîne donnée.
Par exemple : si l'on recherche le motif "pat", le TGen produit sera l'ensemble des mots commençant par "pat" (patriote, pater…).

Pour spécifier la terminaison des mots cherchés, il convient d'utiliser "\>".

---

[6] Pour en savoir plus sur les expressions régulières (xxxxx)

Pour aller plus loin, le site *http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/*

Par exemple, pour rechercher tous les mots qui se terminent par "isme", le motif à utiliser est : "\<.*isme\>". Ce dernier motif peut aussi s'écrire de la manière suivante ".*isme\>", dans la mesure où la recherche se fait sur des mots.

Opérateur	**Fonction**	**Application**
. (le point)	Représente n'importe quel caractère	L'expression "m.l" peut représenter mal, mol…
*	0 ou n occurrences du caractère qui précède	L'expression "com*e" recherche coe, comme, commme,…
+	1 ou n occurrences du caractère qui précèdent	L'expression "com+e" recherche comme, commme,…
\<	Représente un début de mot	L'expression "\<capital" recherche capital, capitale, capitalisme…
\>	Représente une fin de mot	L'expression ".*isme\>" recherche syndicalime, capitalisme…
[ ]	Représente un ensemble de caractères	L'expression "[aeiou]" représente un des caractère de l'ensemble des voyelles. L'expression "[a-z]" représente un des caractères compris entre a et z.
[^]	Représente la négation du contenu de l'ensemble de caractères	L'expression "[^aeiou]" représente des caractères qui ne sont pas ceux de l'ensemble des voyelles.

2.6 Le *Garde-Mots*

Le garde-mots permet de mémoriser formes, segments, *TGens* pour une utilisation ultérieure.
Pour stocker un *TGen* dans le garde-mots il suffit de le faire glisser sur l'icône du cube rouge (cf. glisser/déposer *supra*).
Pour utiliser un *TGen* stocké dans le garde-mots on le glisse à partir du cube rouge jusqu'à la fenêtre de travail (concordance, ventilation des fréquences, carte des sections, etc.) dans laquelle il doit être visualisé.

3 Outils d'analyse statistique

On a regroupé dans ce chapitre plusieurs méthodes qui vont de la description statistique élémentaire (comptages, histogrammes, etc.) à divers types d'analyse multidimensionnelle des données textuelles (analyse factorielles des correspondances, classification automatique, analyse des séries textuelles chronologiques).

3.1 Découpage en parties

Les différentes clés introduites avant la segmentation automatique (cf. section 1 - les corpus de texte) permettent d'opérer différentes partitions du corpus.
Pour réaliser une partition du corpus, on sélectionne un type de clé ; les différents contenus affectés à cette clé découpent alors le corpus en autant de parties différentes.

Exemple : Après avoir segmenté le corpus ***Duchn.txt***, cliquer sur l'icône ***Statistiques par partie***, une boîte de dialogue apparaît, qui permet de choisir une clé de partition (**Figure 3.1**). Sélectionner par exemple la clé *semaine* (double clic ou bouton ***Créer***).

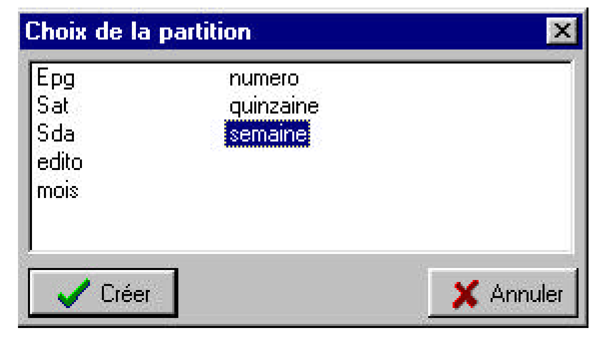


Figure 3.1: Choix d'une partition

Une fenêtre s'ouvre alors qui va permettre de comparer la fréquences des unités textuelles dans l'ensemble des parties.

Ventilation d'une forme (ou Tgen)

En faisant glisser sur cette fenêtre les formes et/ou les segments répétés (section 2.4).qui se trouvent dans les fenêtres à gauche de l'écran, on obtient la ventilation de la ou des unités textuelles choisies, dans les différentes parties du corpus (**Figure 3.2**) . On peut également faire glisser sur cette fenêtre les groupes de formes (section 2.5) réalisés dans la fenêtre correspondante ainsi que les liens stockés dans le *garde-mots* (section 2.6).
On choisit la couleur de traçage du *TGen* à représenter en activant la palette des couleurs située en haut à droite du dictionnaire (resp. de la fenêtre du *groupe de formes*). Si aucune couleur n'est choisie par l'utilisateur, le logiciel sélectionne des couleurs différentes pour chaque nouvelle ventilation.
La zone de traçage peut être réinitialisée à tout moment (bouton ***effacer***, éventuellement après avoir intégré le graphique au rapport).
On peut visualiser la ventilation de plusieurs unités textuelles dans les parties du corpus exprimée :

✍ en fréquence absolues (nombre d'occurrences dans la partie)
✍ en fréquence relatives (nombre d'occurrences rapporté à la longueur de la partie)
✍ en termes de spécificités (résultat d'un calcul statistique, section 3.2).

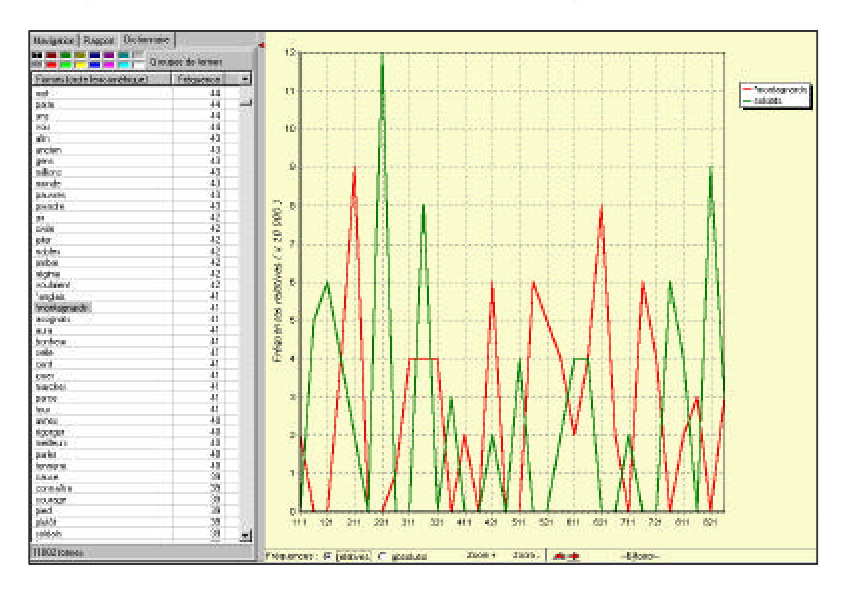


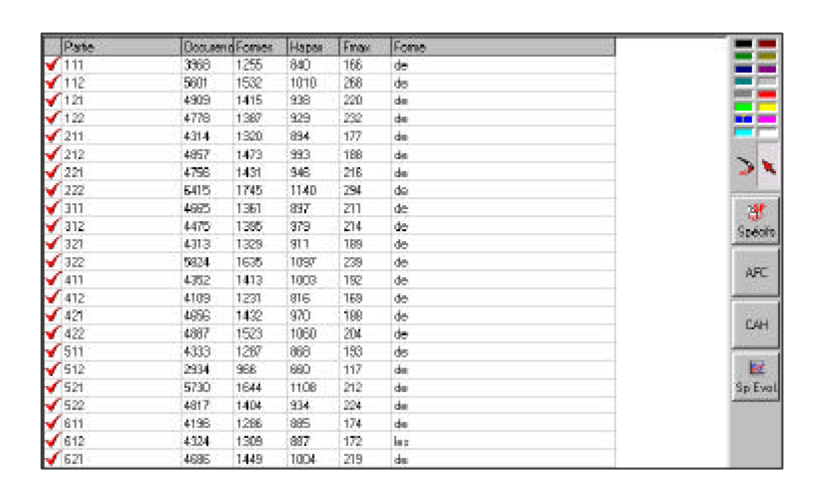
Figure 3.2 : Ventilation d'une forme dans les parties d'un corpus

Statistiques par partie (PCLC)

(principales caractéristiques lexicométriques du corpus et de la partition)

La sélection de l'icône ***PCLC***, fait apparaître les principales caractéristiques par partie suivant la partition choisie.

✍ une coche rouge dans la colonne la plus à gauche indique que la partie est sélectionnée pour le décompte des fréquences globales dans le corpus.
✍ la seconde colonne donne les noms des différentes parties (ici le numéro de la semaine).
✍ la colonne ***occurrences*** indique le nombre des occurrences des formes répertoriées.
✍ La colonne ***formes*** indique le nombre des formes graphiques présentes dans chaque partie.
✍ La colonne ***hapax*** indique, pour chaque partie, le nombre des formes qui n'apparaissent qu'une fois dans la partie.
✍ La colonne ***fréquence maximale*** indique le nombre des occurrences de la forme la plus fréquente.



	Partie	Documen	Formes	Hapax	Fmax	Forme
✓	111	3968	1255	840	166	de
✓	112	5601	1532	1010	268	de
✓	121	4909	1415	938	220	de
✓	122	4778	1387	929	232	de
✓	211	4314	1320	894	177	de
✓	212	4857	1473	993	188	de
✓	221	4798	1431	946	216	de
✓	222	6415	1745	1140	294	de
✓	311	4665	1361	897	211	de
✓	312	4475	1385	979	214	de
✓	321	4313	1329	911	189	de
✓	322	5824	1635	1097	239	de
✓	411	4352	1413	1003	192	de
✓	412	4109	1231	816	169	de
✓	421	4656	1432	970	188	de
✓	422	4887	1523	1060	204	de
✓	511	4333	1287	868	193	de
✓	512	2934	968	660	117	de
✓	521	5730	1644	1108	212	de
✓	522	4817	1404	934	224	de
✓	611	4196	1286	885	174	de
✓	612	4324	1309	887	172	les
✓	621	4686	1449	1004	219	de

Figure 3.3 : Caractéristiques de la partition

Ce tableau permet une comparaison visuelle rapide des parties en fonction de leurs caractéristiques lexicométriques les plus importantes.

3.2 Spécificités

L'analyse des spécificités permet de porter un jugement sur la fréquence de chacune des unités textuelles dans chacune des parties du corpus[7].

Le bouton ***Spécifs*** qui se trouve en haut à droite (**Figure 3.3**) permet d'obtenir le tableau des spécificités d'une partie sélectionnée (**Figure 3.5**) ou d'un ensemble de parties[8].

Par défaut, l'indice de spécificité est calculé pour toute les unités dont la fréquence est supérieure à 10, avec un *seuil de probabilité* fixé à 5 % (une fenêtre ***paramétrage du calcul des spécificités*** apparaît avant le début du calcul qui permet à l'utilisateur de modifier ces paramètres, si besoin).

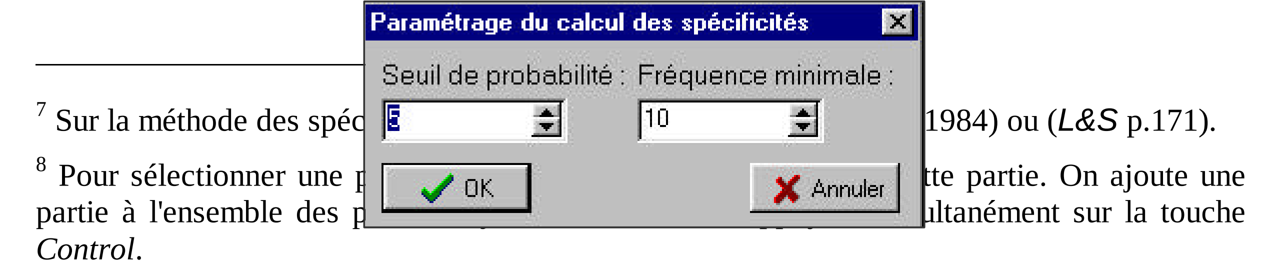
Le *diagnostic de spécificité* calculé contient deux indications.

a) un signe (+ ou –) qui indique un sur-emploi ou un sous-emploi dans la ou les partie(s) sélectionnée(s) par rapport à l'ensemble du corpus.

b) un exposant qui rend compte du degré de significativité de l'écart constaté (un exposant égal à *x*, indique que la probabilité d'un écart de répartition supérieur ou égal à celui que l'on a constaté était, au départ de l'ordre de 10^{-x}).

Exemple : nous F=1270 f= 66 **+05**

indique que la forme *nous*, présente 1270 fois dans le corpus et attestée 66 dans les textes de la semaine numéro 211 est plus fréquente que ce que laissait espérer une répartition "au hasard"[9].



[7] Sur la méthode des spéc… 1984) ou (*L&S* p.171).

[8] Pour sélectionner une p… tte partie. On ajoute une partie à l'ensemble des p… ltanément sur la touche *Control*.

[9] Sous l'hypothèse d'une distribution hypergéométrique avec ces paramètres.

Figure 3.4 : Paramétrage

NB : Si le calcul des segments répétés a été préalablement effectué, les segments spécifiques apparaissent également dans la liste des unités spécifiques.

Résultats du calcul des spécificités

Spécifs - Part : semaine

Corpus de référence : 111 112 121 122 211 212 221 222
Parties sélectionnées : 211
Spécificités (•) positives ( ) négatives

Terme	Frq Tot.	Frq P...	Spécif
nous	1270	66	5
faire	412	25	4
chaque	33	6	4
toi	88	10	4
marcher	41	6	3
année	14	3	3
comment	33	5	3
pendant	77	7	3
millions	43	5	3
présent	24	4	3
savons	12	3	3
mille	55	6	3
vous	1097	52	3
subsistances	47	6	3
département	14	3	3
gueule	12	3	3
avez	171	12	3
aurons	21	4	3
ensuite	22	3	2
*europe	22	3	2
publique	23	3	2
*louis	20	3	2
bled	16	3	2
canon	19	3	2
*st	23	3	2
comités	12	2	2
oeuvre	11	2	2
sol	11	2	2
moutons	11	2	2
noire	11	2	2
ruine	10	2	2
sucre	10	2	2
anglais	10	2	2

Dans la première colonne on trouve les unités spécifiques classées par ordre décroissant de spécificité. Les deux colonnes suivantes indiquent respectivement la fréquence totale de la forme dans l'ensemble du corpus et la fréquence de la forme dans la partie sélectionnée
Les boutons ***positives*** et ***négatives*** de l'onglet des spécificités permettent d'inverser l'ordre de présentation de la liste qui s'ouvre par défaut sur les spécificités positives.

3.3 Spécificités chronologiques

Pour les séries textuelles chronologiques (série de textes produits par une même source textuelle et régulièrement espacés dans le temps, exemple ***Duchesne***), à côté de l'analyse des spécificités de chacune des parties du corpus, l'analyse des spécificités chronologiques met en évidence le vocabulaire particulier de périodes plus larges formées de parties consécutives (cf L&S p197 et Salem 93).

Accroissements spécifiques

Pour une partie sélectionnée, le bouton ***SpEvol***, permet de calculer les spécificités (ou ***accroissements spécifiques***) de cette partie par rapport à l'ensemble des périodes précédentes (en excluant momentanément du corpus les périodes postérieures). Le résultat de ces calculs est fourni sous la forme d'un tableau de spécificités identique à celui présenté à la **Figure 3.5**.
NB : La partie négative des accroissements spécifiques met en évidence des unités textuelles qui ont tendance à être sous-utilisées dans la période considérée par rapport aux périodes qui précèdent.

3.4 Analyse Factorielle des Correspondances (AFC)

Le bouton ***AFC*** permet de réaliser une *analyse factorielle des correspondances* sur l'ensemble des parties du corpus (à l'exclusion de celles qui ont été écartées par suppression de la coche rouge)[10].

La fenêtre de paramétrage (**Figure 3.6**) permet de fixer entre autres :

- Le nombre des unités textuelles prises en compte dans l'analyse
- Le nombre des facteurs à extraire

NB : Par défaut, l'analyse prend en compte les unités dont la fréquence est supérieure à 10. La modification du seuil de fréquence minimale entraîne un nouveau calcul du nombre des unités prises en compte.



Figure 3.6 Le paramétrage de l'AFC:

On lance l'analyse en cliquant sur le bouton ***OK***. Les parties du corpus apparaissent sur le plan des deux premiers axes factoriels extraits par l'analyse. On peut obtenir d'autres visualisations en sélectionnant d'autres axes (boîtes situées au-dessus du graphique factoriel).

Les différents plans factoriels permettent une estimation des proximités calculées entre les différentes parties sélectionnées, en fonction de leur vocabulaire.

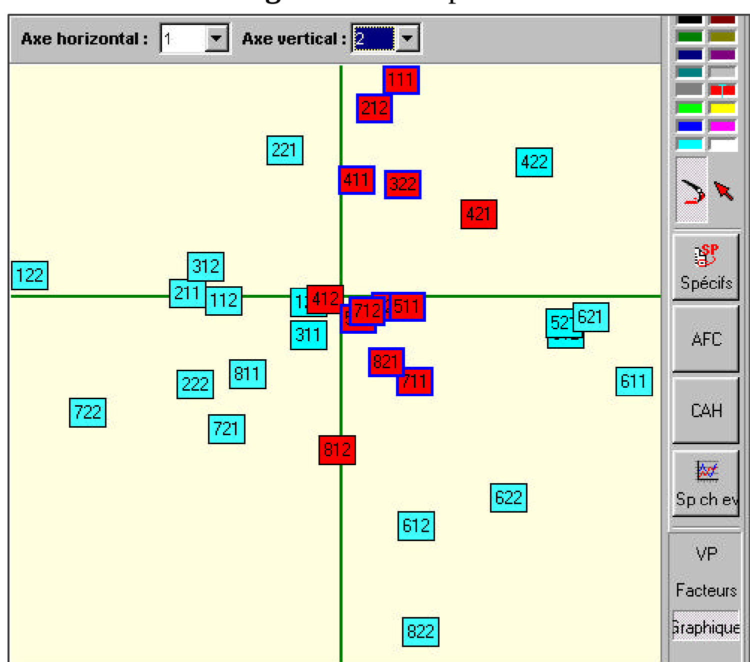
On peut réitérer l'analyse en en écartant certaines parties (clic droit - les parties écartées du corpus apparaissent alors avec des rayures grises).

On peut sélectionner (clic gauche), directement sur la carte produite, une partie ou un groupe de parties. Les contours des parties sélectionnées apparaissent alors en surbrillance. Cette sélection permet, par exemple, de calculer des spécificités sur un groupe de parties.

---

[10] On trouvera un exposé complet sur cette méthode, par exemple, dans (*L&S* p 135).

Figure 3.7 : Graphe AFC



Le pinceau et la boîte de couleurs situés à droite du graphique permettent d'associer une couleur à un ensemble de parties. L'outil flèche permet de passer à nouveau en mode de sélection.

Le dernier groupe de boutons permet de naviguer parmi les résultats de l'analyse.

- ✍ ***VP*** permet de consulter l' *histogramme des valeurs propres*
- ✍ ***Facteurs*** permet de consulter le tableau des facteurs
- ✍ ***Graphique*** permet de revenir au plan factoriel.

4 Outils de navigation lexicométrique

Cette section décrit des fonctionnalités qui permettent de se déplacer parmi les résultats produits par les différentes méthodes lexicométriques et le texte initial.

4.1 Carte des sections

La ***carte des sections*** permet une visualisation du corpus découpé en sections par la promotion d'un (ou de plusieurs) caractère particulier (paragraphes, point, etc.) au statut de ***délimiteur de section***.

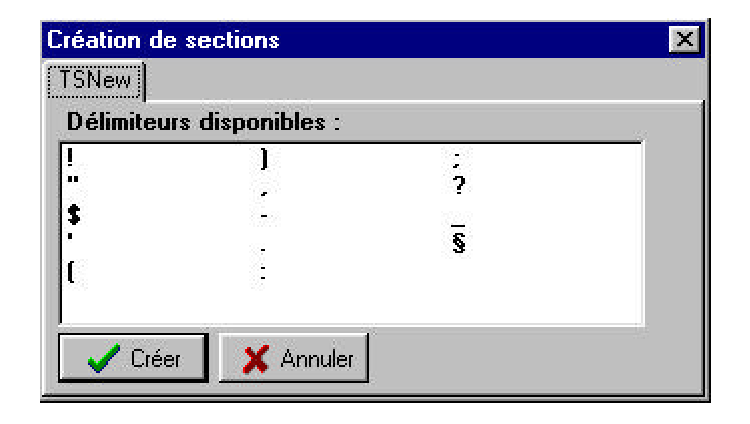


Figure 4.1 : Choix des délimiteurs de section

Faire une carte des sections pour un Tgen

Sélectionner le *Tgen* (à partir du dictionnaire, du *Garde-mots*, de la liste des segments répétés, etc.…) et le faire glisser sur la carte (bouton gauche maintenu enfoncé).

- On sélectionne la section à visualiser dans la fenêtre du bas en cliquant sur le carré qui la représente dans la carte des sections.
- On agrandit la taille des carrés qui représente chacune des sections en déplaçant vers la droite le curseur situé en haut et à gauche de la fenêtre.
- On matérialise une partition activée en la sélectionnant dans la boite de liste située immédiatement à la droite de ce curseur.
- On colorie les sections en fonction de la spécificité du *Tgen* étudié, dans la section. On coche d'abord la case ***seuil***. L'icône qui précède immédiatement permet de régler deux seuils en probabilités qui entraîneront un coloriage (plus ou moins sombre) des sections. Pour une représentation simultanée de deux *Tgens*, ce processus peut être réitéré (en prenant soin de changer la couleur dans la boite correspondante). Il faut maintenir, dans ce cas, la touche *Control* en position basse lors du second *glisser/déposer*.

Les outils statistiques de la carte des sections

Les deux icônes situées au même niveau à droite de la fenêtre permettent de repérer les types caractéristiques d'un ensemble de sections (spécificités des sections sélectionnées, cf. 3.2)

✍ Le premier bouton ***Cooccurences*** constitue automatiquement une sélection des sections dans lesquelles le *Tgen* étudié est présent (c'est cet ensemble de sections que l'on compare à l'ensemble du corpus).

✍ Le deuxième bouton ***Spécificités*** permet à l'utilisateur de constituer une sélection arbitraire de sections dont on étudiera ensuite le vocabulaire spécifique (selon les conventions *Windows*, on sélectionne les sections une à une en maintenant le bouton ***Control*** en position basse ; la touche majuscule permet de sélectionner un groupe de sections consécutives).

Comme toujours, les listes de spécificités sont affichées dans la fenêtre de gauche. Le nombre des sections concernées par la sélection apparaît en haut de la fenêtre ; un bouton ***ajouter au rapport Section*** placé en bas de la fenêtre permet de sauvegarder les résultats.

Naviguer à l'aide de la carte des sections



✍ Les boutons situés à gauche de la fenêtre de visualisation de la sélection (en forme de mains) permettent de passer, respectivement, à la section suivante/précédente ou à l'occurrence suivante/précédente du *Tgen* sélectionné.

✍ L'icône *Ajouter au rapport section* permet d'enregistrer la section visualisée dans la fenêtre du bas.

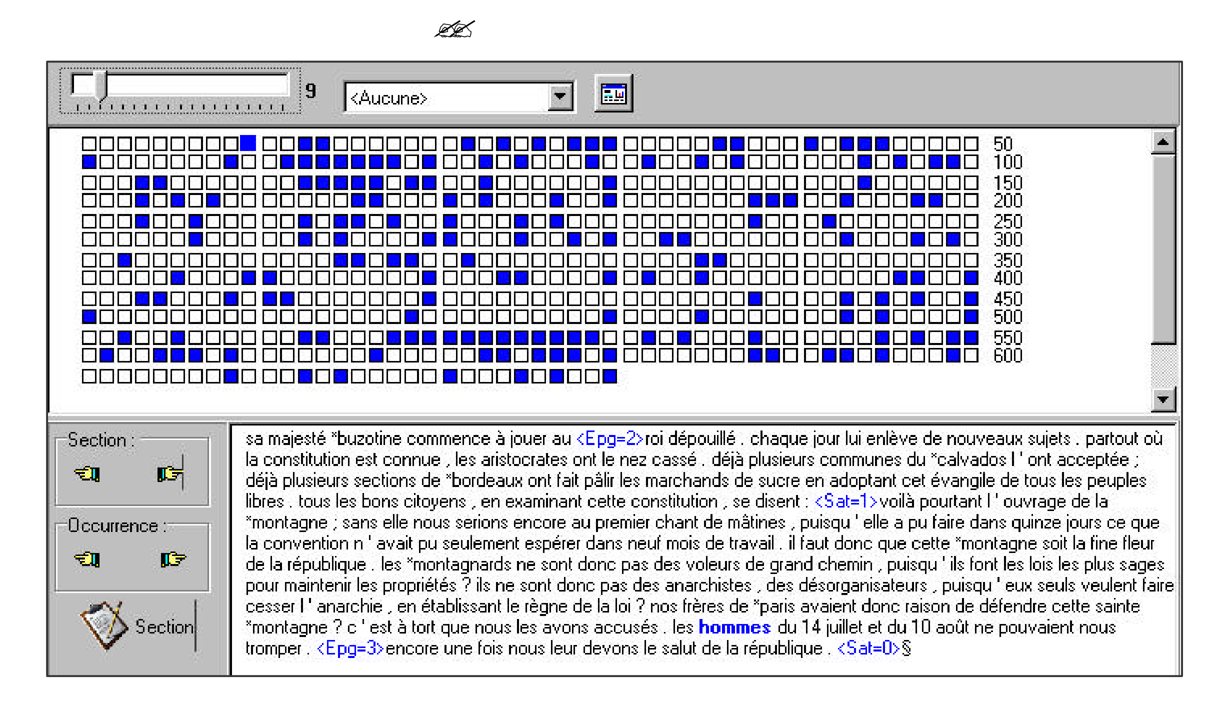


Figure 4.2 : Ventilation dans les paragraphes de la forme *hommes*

4.2 Mieux utiliser les fenêtres de travail

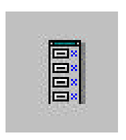
Créer une feuille de travail

Pour éviter un fractionnement trop important de la fenêtre de travail principale, il est préférable de créer de nouvelles feuilles de travail en cliquant sur cette icône.
Les feuilles de travail s'empilent sur la droite de la fenêtre principale. Les onglets "Feuille n°i" permettent de passer de l'une à l'autre. On peut transporter des liens *Tgen* d'une feuille à l'autre en passant par exemple par le *Garde-mot*.

Déplacer vers une autre feuille de travail

Pour déplacer une fenêtre-résultat vers une nouvelle feuille, la sélectionner, cliquer sur l'icône puis sélectionner la feuille désirée.

Nouvelle feuille
Feuille 2
Feuille 3
Feuille 4
Feuille 5
Feuille 6



Mosaïque

Cette icône permet de réorganiser plusieurs fenêtres sur la même feuille.

4.3 Le rapport



Le dossier ***Rapport*** contient les résultats sélectionnés par l'utilisateur pour une exploitation ultérieure. Ce dossier aisément manipulable à l'aide d'un navigateur web (*Internet Explorer* , *Netscape*, etc.) contient un fichier *index.htm* qui permet la navigation parmi les résultats.
Le rapport peut être consulté à tout moment à la condition que l'utilisateur l'ait préalablement enregistré (bouton ***Enregistrer*** au bas de l'onglet ***Rapport***).).

Editer les résultats

Pour visualiser un texte ou bien les résultats obtenus à partir de ***Lexico 3***, cliquer sur l'icône "Editeur" et à partir de l'icône "Ouvrir" sélectionner le document désiré.
Pour conserver les documents stockés lors de sessions différentes , il est préférable de sauvegarder chaque fois le dossier ***Rapport*** dans un dossier (ou sous un nom) différent.
On trouve le dossier ***Rapport*** dans le dossier ***Lexico3*** créé par l'installation du logiciel.

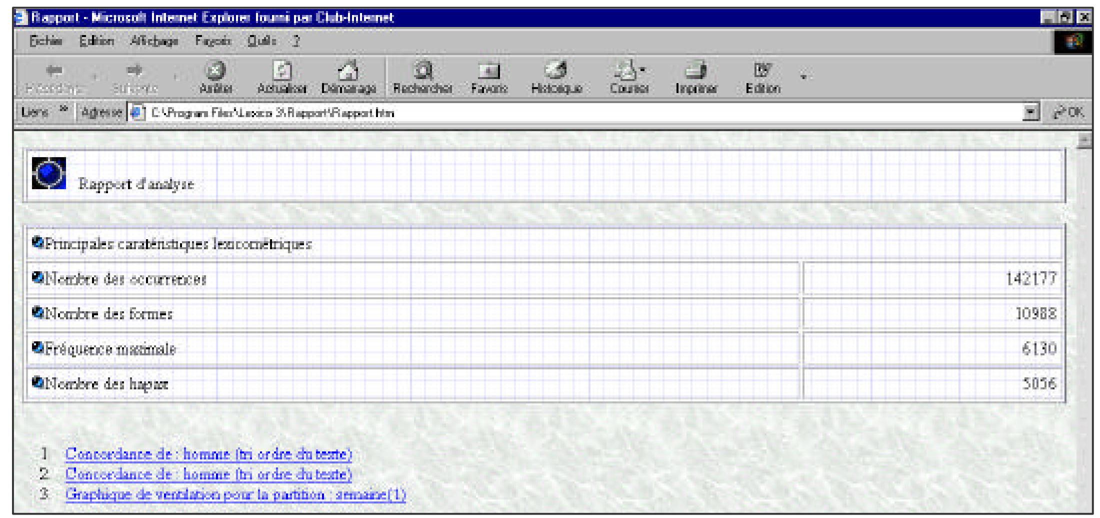


Figure 4.6 : Rapport

4.4.Options- Aides - Compléments

Options

Ce bouton permet de modifier les limites du logiciel (100 000 formes lexicales différentes environ) lors du traitement de gros corpus (plusieurs millions d'occurrences). Il permet aussi d'indiquer si le corpus traité a été préalablement étiqueté.

Quelques exemples de corpus :

Corpus	*pages*	*occurrences*	*formes différentes*	*fréquence max.*
Duchesne	350	142 177	10 988	6130 (*de*).
Coran (trad. Fr)				(*de*).

Duchesne	(*de*).

Onglet de navigation

Cet onglet permet de naviguer parmi les résultats produits par ***Lexico3*** de la même manière que l'explorateur Windows.

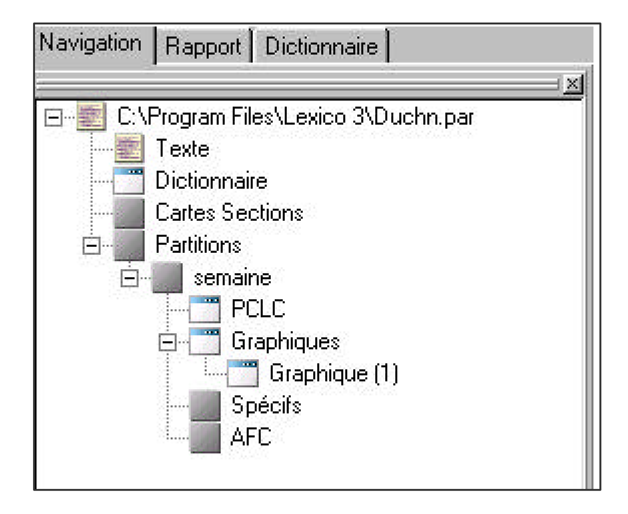
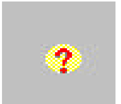


Figure 4.6 : Navigation

Plein écran

Pour visualiser la fenêtre de droite en plein écran, cliquer sur la flèche rouge située entre les fenêtres gauche et droite.



Aide

Le fichier d'aide de ***Lexico3*** (qui contient le présent manuel) peut être consulté à tout moment à partir de la console en cliquant sur l'icône ***Aide***.

Quitter

Avant de quitter ***Lexico3***, vérifier que toutes les données sont bien sauvegardées dans le rapport, puis cliquer sur l'icône.

5 Glossaire pour la statistique textuelle

La définition de quelques notions de base en statistique textuelle est reprise dans l'aide en ligne.

NB : Les astérisques renvoient à une entrée de ce même glossaire. Les abréviations qui suivent entre parenthèses précisent le domaine auquel s'applique plus particulièrement la définition.

Abréviations :
ac Analyse factorielle des correspondances
acm Analyse des correspondances multiples
cla Classification
sp Méthode des Spécificités
sr Analyse des segments répétés
ling Linguistique
stat Statistique
sa Segmentation automatique

accroissement spécifique - (sp) spécificité* calculée pour une partie d'un corpus par rapport à une partie antérieure

analyse factorielle (stat) - famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

analyse des correspondances (stat)- méthode d'analyse factorielle s'appliquant à l'étude de tableaux à double entrée composés de nombres positifs. L'AC est caractérisée par l'emploi d'une distance (ou métrique) particulière dite distance du chi-2 (ou c2).

caractère (sa) - signe typographique utilisé pour l'encodage du texte sur un support lisible par l'ordinateur.

caractères délimiteurs / non-délimiteurs (sa) - distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences* (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.

- les caractères **délimiteurs de séquences** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.

- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

classification (stat) - technique statistique permettant de regrouper des observations ou des individus entre lesquels a été définie une distance.

classification hiérarchique (cla) - technique particulière de classification produisant par agglomération progressive des classes ayant la propriété d'être, pour deux quelconques d'entre-elles, soit disjointes, soit incluses.

concordance (sa) - l'ensemble de lignes de contexte se rapportant à une même forme-pôle.

contribution absolue (ou contribution) - (ac) contribution apportée par un élément au facteur. Pour un facteur donné, la somme des contributions sur les éléments de chacun des ensembles mis en correspondance est égale à 100.

contribution relative (ou cosinus carré) - (ac) contribution apportée par le facteur à un élément. Pour un élément donné, la somme des contributions relatives sur l'ensemble des facteurs est égale à 1.

cooccurrence (sa) - (une c.) - présence simultanée, mais non forcément contiguë, dans un fragment de texte (séquence, phrase, paragraphe, voisinage d'une occurrence, partie du corpus etc.) des occurrences de deux formes données.

corpus (ling) - ensemble limité des éléments (énoncés) sur lesquels se base l'étude d'un phénomène linguistique.

(lexicométrie) ensemble de textes réunis à des fins de comparaison; servant de base à une étude quantitative.

délimiteurs de séquence - (sa) sous-ensemble des caractères délimiteurs* de forme* correspondant aux ponctuations faibles et fortes (en général - le point, le point d'interrogation, le point d'exclamation, la virgule, le point-virgule, les deux points, les guillemets, les tirets et les parenthèses).

dendrogramme - (cla) représentation graphique d'un arbre de classification hiérarchique, mettant en évidence l'inclusion progressive des classes.

discours/langue - La langue est un ensemble virtuel qui ne peut être appréhendé que dans son actualisation orale ou écrite; "discours" est un terme commode qui recouvre les deux domaines de cette actualisation.

distance du chi-2 - distance entre profils* de fréquence utilisée en analyse des correspondances* et dans certains algorithmes* de classification*.

éditions de contextes (sa) - éditions de type concordanciel dans lesquelles les occurrences d'une forme sont accompagnées d'un fragment de contexte pouvant contenir plusieurs lignes de texte autour de la forme-pôle. La longueur de ce contexte est définie en nombre d'occurrences avant et après chaque occurrence de la forme-pôle.

éléments d'un segment (sr) - chacune des formes correspondant aux occurrences qui entrent dans sa composition. ex : A, B, C sont respectivement les premier, deuxième et troisième éléments du segment ABC.

éléments actifs- (ac ou acm) ensemble des éléments servant de base au calcul des axes factoriels, des valeurs propres relatives à ces axes et des coordonnées factorielles.

éléments supplémentaires (ou illustratifs)- (ac ou acm) ensemble des éléments ne participant pas aux calculs des axes factoriels, pour lesquels on calcule des coordonnées factorielles qui auraient été affectées à une forme ayant la même répartition dans le corpus mais participant à l'analyse avec un poids négligeable.

énoncé/énonciation - (ling) à l'intérieur du texte un ensemble de traces qui manifestent l'acte par lequel un auteur a produit ce texte.

facteur- (ac ou acm) variables artificielles construites par les techniques d'analyse factorielle permettant de résumer (de décrire brièvement) les variables actives initiales.

forme- (sa) ou "**forme graphique**" archétype correspondant aux occurrences* identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence.

forme banale - (sp) pour une partie du corpus donnée, forme ne présentant aucune spécificité (ni positive ni négative) dans cette partie .

forme caractéristique - (d'une partie) synonyme de spécificité positive*.

forme commune - forme attestée dans chacune des parties du corpus.

forme originale- (pour une partie du corpus) forme trouvant toutes ses occurrences dans cette seule partie.

fréquence (sa) - (d'une unité textuelle) le nombre de ses occurrences dans le corpus.

fréquence d'un segment (sr) - (ou d'une polyforme) le nombre des occurrences de ce segment, dans l'ensemble du corpus.

fréquence maximale (sa) - fréquence de la forme la plus fréquente du corpus (en français, le plus souvent, la préposition "de").

fréquence relative (sa) - la fréquence d'une unité textuelle dans le corpus ou dans l'une de ses parties, rapportée à la taille du corpus (resp. de cette partie).

gamme des fréquences (sa) **-** suite notée Vk, des effectifs correspondant aux formes de fréquence k, lorsque k varie de 1 à la fréquence maximale.

hapax - gr. hapax (legomenon), "chose dite une seule fois".

(sa) forme dont la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie).

identification - (stat, ling, sa) reconnaissance d'un seul et même élément à travers ses multiples emplois dans des contextes et dans des situations différentes.

index - (sa) liste imprimée constituée à partir d'une réorganisation des formes et des occurrences d'un texte, ayant pour base la forme graphique et permettant de regrouper les références* relatives à l'ensemble des occurrences d'une même forme.

index alphabétique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicographique* (celui des dictionnaires).

index hiérarchique (sa) - index* dans lequel les formes-pôles* sont classées selon l'ordre lexicométrique*.

index par parties - ensemble d'index (hiérarchiques ou alphabétiques) réalisés séparément pour chaque partie d'un corpus.

lemmatisation - regroupement sous une forme canonique (en général à partir d'un dictionnaire) des occurrences du texte. En français, ce regroupement se pratique en général de la manière suivante :

- les formes verbales à l'infinitif,
- les substantifs au singulier,
- les adjectifs au masculin singulier,
- les formes élidées à la forme sans élision.

lexical - (ling) qui concerne le lexique* ou le vocabulaire*.

lexicométrie ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire* d'un corpus de textes.

lexique - (ling) ensemble virtuel des mots d'une langue.

longueur (sa) - (d'un corpus, d'une partie de ce corpus, d'un fragment de texte, d'une tranche, d'un segment, etc.) le nombre des occurrences contenues dans ce corpus (resp. : partie, fragment, etc.). Synonyme : *taille*.

On note: T la longueur du corpus; t j celle de la partie (ou tranche) numéro j du corpus.

longueur d'un segment (sr) - le nombre des occurrences entrant dans la composition de ce segment.

occurrence (sa) **-** suite de caractères non-délimiteurs bornée à ses extrémités par deux caractères délimiteurs* de forme.

ordre lexicographique -

_ pour les formes graphiques :

l'ordre selon lequel les formes sont classées dans un dictionnaire.

NB : Les lettres comportant des signes diacrisés sont classées au même niveau que les mêmes caractères non diacrisés, le signe diacritique n'intervenant que dans les cas d'homographie complète. Dans les dictionnaires, on trouve par exemple rangées dans cet ordre les formes : *mais, maïs, maison, maître* .

_ pour les polyformes:

ordre résultant d'un tri des polyformes par ordre lexicographique sur la première composante. Les polyformes commençant par une même forme graphique sont départagées par un tri lexicographique sur la seconde, etc.

ordre lexicométrique (sa) -

_ pour les formes graphiques :

ordre résultant d'un tri des formes du corpus par ordre de fréquences décroissantes ; les formes de même fréquence sont classées par ordre lexicographique.

_ pour les polyformes:

ordre résultant d'un tri par ordre de longueur décroissante des segments, les segments de même longueur sont départagés par leur fréquence, les segments ayant même longueur et même fréquence par l'ordre lexicographique.

paradigme- (ling) ensemble des termes qui peuvent figurer en un point de la chaîne parlée.

paradigmatique- (sa) qui concerne le regroupement en série des unités textuelles, indépendamment de leur ordre de succession dans la chaîne écrite.

partie - (d'un corpus de textes) fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

partition - (d'un corpus de textes) division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

(d'un ensemble, d'un échantillon) division d'un ensemble d'individus ou d'observations en *classes* disjointes dont la réunion est égale à l'ensemble tout entier.

partition longitudinale - (sa) partition d'un corpus en fonction d'une variable qui définit un ordre sur l'ensemble des parties

périodisation (sa) - regroupement des parties naturelles du corpus respectant l'ordre chronologique d'écriture, d'édition ou de parution des textes réunis dans le corpus.

phrase - (sa) fragment de texte compris entre deux séparateurs* de phrase.

polyforme (sr) - archétype des occurrences d'un segment; suite de formes non séparées par un séparateur de séquence, qui n'est pas obligatoirement attestée dans le corpus.

ponctuation - Système de signes servant à indiquer les divisions d'un texte et à noter certains rapports syntaxiques et/ou conditions d'énonciation.

(*sa*) caractère (ou suite de caractères) correspondant à un signe de ponctuation.

pourcentages d'inertie - (ac ou acm) quantités proportionnelles aux valeurs propres* dont la somme est égale à 100. Notées ta.

profil - (stat et ac) (d'une ligne ou d'une colonne d'un tableau à double entrée) vecteur constitué par le rapport des effectifs contenus sur cette ligne (resp. colonne) à la somme des effectifs que contient la ligne (resp. la colonne).

répartition (sa) - (des occurrences d'une forme dans les parties du corpus) nombre des parties du corpus dans lesquelles cette forme est attestée.

section - (sr) portion de texte comprise entre deux délimiteurs de section (exemple : le paragraphe, etc.).

segment - (sr) toute suite d'occurrences consécutives dans le corpus et non séparées par un séparateur* de séquence est un segment du texte.

segment répété (sr) - (ou polyforme répétée) suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

segmentaire - (sr) ensemble des termes* attestés dans le corpus.

segmentation - opération qui consiste à délimiter des unités minimales* dans un texte.

segmentation automatique - ensemble d'opérations réalisées au moyen de procédures informatisées qui aboutissent à découper, selon des règles prédéfinies, un texte stocké sur un support lisible par un ordinateur en unités distinctes que l'on appelle des unités minimales*.

séparateurs de phrases - (sa) sous-ensemble des caractères délimiteurs* de séquence* correspondant aux seules ponctuations fortes (en général : le point, le point d'interrogation, le point d'exclamation).

séquence - (sa) suite d'occurrences du texte non séparées par un délimiteur* de séquence.

seuil - (stat) quantité arbitrairement fixée au début d'une expérience visant à sélectionner parmi un grand nombre de résultats, ceux pour lesquels les valeurs d'un indice numérique dépassent ce seuil (de fréquence, en probabilité, etc.).

sous-fréquence (sa) - (d'une unité textuelle dans une partie, tranche, etc.) nombre des occurrences de cette unité dans la seule partie (resp. tranche, etc.) du corpus.

sous-segments (sr) - pour un segment donné, tous les segments de longueur inférieure et compris dans ce segment sont des sous-segments. ex : AB et BC sont deux sous-segments du segment ABC.

spécificité chronologique - (sp) spécificité* portant sur un groupe connexe de parties d'un corpus muni d'une partition longitudinale*.

spécificité positive - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

spécificité négative - (sp) pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ.

stock distributionnel du vocabulaire - (d'un fragment de texte) le vocabulaire* de ce fragment assorti de comptages de fréquence pour chacune des formes entrant dans sa composition.

syntagmatique- (sa) qui concerne le regroupement des unités textuelles, selon leur ordre de succession dans la chaîne écrite.

syntagme- (ling) groupe de mots en séquence formant une unité à l'intérieur de la phrase.

tableau de contingence (stat) - synonyme de tableau de fréquences ou de tableau croisé: tableau dont les lignes et les colonnes représentent respectivement les modalités de deux questions (ou deux variables nominales) , et dont le terme général représente le nombre d'individus correspondant à chaque couple de modalités.

tableau lexical entier (TLE) - tableau à double entrée dont les lignes sont constituées par les ventilations* des différentes formes dans les parties du corpus. Le terme générique *k(i,j)* du TLE est égal au nombre de fois que la forme *i* est attestée dans la partie *j* du corpus. Les lignes du TLE sont triées selon l'ordre lexicométrique* des formes correspondantes.

tableau des segments répétés (TSR) - tableau à double entrée dont les lignes sont constituées par les ventilations* des segments répétés dans les parties du corpus. Les lignes du TSR sont triées selon l'ordre lexicométrique* des segments. (i.e. longueur décroissante, fréquence décroissante, ordre lexicographique).

tableau lexical- tableau à double entrée résultant du TLE par suppression de certaines lignes (par exemple celles qui correspondent à des formes dont la fréquence est inférieure à un seuil donné).

taille- (sa) (d'un corpus) sa longueur* mesurée en occurrences (de formes simples).

terme - (sr) nom générique s'appliquant à la fois aux formes* et aux polyformes*. Dans le premier cas on parlera de termes de longueur 1. Les polyformes sont des termes de longueur 2,3, etc.

termes contraints / termes libres - Un terme S1 est contraint dans un autre terme S2 de longueur supérieure si toutes ses occurrences* sont des sous-segments* de segments correspondant à des occurrences du segment S2. Si au contraire un terme possède plusieurs expansions distinctes, qui ne sont pas forcément récurrentes, c'est un terme libre.

types généralisés (Tgens)- unités de dépouillement définies par l'utilisateur à l'aide d'outils permettant d'effectuer automatiquement des regroupements d'occurrences du texte (ex : les occurrences des formes qui commencent par la séquence de caractère *patr : patrie, patriotes, patriotisme, etc.*).

unités minimales (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent)

valeur modale - (stat) valeur pour laquelle une distribution atteint son maximum.

valeurs propres - (ac ou acm) quantités permettant de juger de l'importance des facteurs successifs de la décomposition factorielle. La valeur propre notée la. mesure la dispersion des éléments sur l' axe.a.

valeurs-tests - (ac ou acm) quantités permettant d'apprécier la signification de la position d'un élément supplémentaire* (ou illustratif) sur une axe factoriel. Brièvement, si une valeur test dépasse 2 en valeur absolue, il y a 95 chances sur 100 que la position de l'élément correspondant ne puisse être due au hasard.

variables actives - variables utilisées pour dresser une typologie, soit par analyse factorielle, soit par classification. Les typologies dépendent du choix et des poids des variables actives, qui doivent de ce fait constituer un ensemble homogène.

variables supplémentaires (ou illustratives) - variables utilisées *a posteriori* pour illustrer des plans factoriels ou des classes. Une variable supplémentaire peut-être considérée comme une variable active munie d'un poids nul.

variables de type T - variable dont la fréquence est à peu près proportionnelle à l'allongement du texte. (ex : la fréquence maximale)

variables de type V- variable dont l'accroissement a tendance à diminuer avec l'allongement du texte (ex : le nombre des formes, le nombre des hapax).

ventilation (sa) - (des occurrences d'une unité dans les parties du corpus) La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences* de cette unité dans chacune des parties, prises dans l'ordre des parties.

vocabulaire (sa) - ensemble des formes* attestées dans un corpus de textes.

vocabulaire commun - (sa) l'ensemble des formes attestées dans chacune des parties du corpus.

vocabulaire de base - (sp) ensemble des formes du corpus ne présentant, pour un seuil fixé, aucune spécificité (négative ou positive) dans aucune des parties , (i.e. l'ensemble des formes qui sont "banales" pour chacune des parties du corpus).

vocabulaire original- (sa) (pour une partie du corpus) l'ensemble des formes* originales* pour cette partie.

voisinage d'une occurrence - (sa) pour une occurrence donnée du texte, tout segment (suite d'occurrences consécutives, non séparées par un délimiteur de séquence) contenant cette occurrence.

Références bibliographiques

Baayen H. (2001) - “Word Frequency Distributions “, *Series: Text, Speech and Language Technology*, Volume 18, Kluwer Academic Publishers, Dordrecht Hardbound.

Bécue M. (1988) - Characteristic repeated segments and chains in textual data analysis, COMPSTAT, 8th Symposium on Computational Statistics, Physica Verlag, Vienna.

Becue M., Peiro R. (1993) - Les quasi-segments pour une classification automatique des réponses ouvertes, in Actes des 2ndes Journées Internationales d'analyse des données textuelles, (Montpellier), ENST, Paris, p 310-325.

Benzécri J.-P. & coll. (1973) - La taxinomie, Vol. I ; L'analyse des correspondances, Vol. II, Dunod, Paris.

Benzécri J.-P. (1991a) - Typologies de textes grecs d'après les occurrences des formes des mots-outil, Les Cahiers de l'Analyse des Données, XVI, n°1, p 61-86.

Benzécri J.-P.& coll. (1981a) - Pratique de l'analyse des données, tome 3, Linguistique & Lexicologie, Dunod , Paris.

Bernet C. (1983) - Le vocabulaire des tragédies de Jean Racine, Analyse statistique, Slatkine-Champion, Genève 1983.

Biber D., Conrad S., Reppen R. (1998) - *Corpus Linguistics : Investigating language structure and use*, Cambridge University Press.

Bolasco S. (1992) - Sur différentes stratégie dans une analyse des formes textuelles : Une expérimentation à partir de données d'enquête, Jornades Internacionals d'Analisi de Dades Textuals, UPC, Barcelona, p 69-88.

Bonnafous S. (1991) - L'immigration prise aux mots. Les immigrés dans la presse au tournant des années quatre-vingt, Kimé, Paris.

Bouillon P. (1998), - *Traitement automatique du langage naturel*, Editions Duculot.

Brunet E. (1981) - Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française, Slatkine-Champion, Genève-Paris.

Crochemore M., Hancart C., Lecroq T. (2001) - *Algorithme du texte*, Vuibert.

Demonet M., Geffroy A., Gouaze J., Lafon P., Mouillaud M., Tournier M. (1975) - Des tracts en Mai 68. Mesures de vocabulaire et de contenu, Armand Colin et Presses de la Fondation Nat. des Sc. Pol., Paris.

Dendien J. (1986) - La Base de données de l'Institut National de la Langue Française, Actes du colloque international CNRS, Nice, juin 1985, 2 vol., Slatkine-Champion Genève, Paris.

Desgraupes B. (2001) *Introduction aux expressions régulières* , Vuibert.

Geffroy A., Lafon P., Tournier M. (1974) - L'indexation minimale, Plaidoyer pour une non-lemmatisation, Colloque sur l'analyse des corpus linguistiques : "Problèmes et méthodes de l'indexation minimale", Strasbourg 21-23 mai 1973.

Gobin C., Deroubaix J. C. (1987) - Du progrès, de la réforme de l'Etat, de l'austérité. Déclarations gouvernementales en Belgique, Mots, n°15, p 137-170.

Guilbaud G.-Th. (1980) - Zipf et les fréquences, Mots N° 1, p 97-126.

Guilhaumou J. (1986) - L'historien du discours et la lexicométrie. Etude d'une série chronologique : Le père Duchesne de Hébert, juillet 1793- mars 1794, Histoire & Mesure , Vol. I, n° 3-4.

Guiraud P. (1954) - Les caractères statistiques du vocabulaire, P.U.F., Paris.

Guiraud P. (1960) - Problèmes et méthodes de la statistique linguistique, P.U.F., Paris.

Guttman L. (1941) - The quantification of a class of attributes: a theory and method of a scale construction, in The prediction of personal adjustment (P. Horst, ed.), SSCR New York, p 251 -264.

Habert B., Fabre C., Issac F. (1998) - *De l'écrit au numérique (constituer, normaliser et exploiter les corpus électroniques)*, InterEditions.

Habert B., Salem A., Nazarenko A. (1997) - *Les linguistiques de corpus*, Armand Colin, Paris.

Habert B., Tournier M. (1987) - La tradition chrétienne du syndicalisme français aux prises avec le temps. Evolution comparée des résolutions confédérales (1945 - 1985), Mots, n°14.

Jurafsky D., Martin J. H. (2000) - "Speech and Language Processing : An Introduction to Natural Language Processing", *Computational Linguistics, and Speech Recognition,* Prentice-Hall.

Labbé D. (1983) - François Mitterrand - Essai sur le discours, La pensée sauvage, Grenoble.

Labbé D. (1990) - Le vocabulaire de François Mitterrand, Presses de la Fond. Nat. des Sciences Politiques, Paris.

Labbé D. (1990) - Normes de dépouillement et procédures d'analyse des textes politiques, CERAT, Grenoble.

Labbé D., Thoiron P., Serant D. (Ed.) (1988) - Etudes sur la richesse et la structure lexicales, Slatkine-Champion, Paris-Genève.

Lafon P. (1980) - Sur la variabilité de la fréquence des formes dans un corpus, Mots N°1 , p 127-165.

Lafon P. (1981) - Analyse lexicométrique et recherche des cooccurrences, Mots N°3 , p 95-148.

Lafon P. (1981) - Dépouillements et statistiques en lexicométrie, Slatkine-Champion, 1984, Paris.

Lafon P., Salem A. (1983) - L'Inventaire des segments répétés d'un texte, Mots N°6, p 161-177.

Lafon P., Salem A., Tournier M. (1985) - Lexicométrie et associations syntagmatiques (Analyse des segments répétés et des cooccurrences appliquée à un corpus de textes syndicaux). Colloque de l'ALLC, Metz -1983, Slatkine-Champion, Genève, Paris, p 59-72.

Lebart L. (1969) - L'Analyse statistique de la contiguïté, Publications de l'ISUP, XVIII- p 81 - 112.

Lebart L. (1982b) - L'Analyse statistique des réponses libres dans les enquêtes socio-économiques, Consommation, n°1, Dunod, p 39-62.

Lebart L., Salem A. (1988) - Analyse statistique des données textuelles, Dunod, Paris.

Lebart L., Salem A., Berry E. (1991) - Recent development in the statistical processing of textual data, Applied Stoch. Model and Data Analysis, 7, p 47-62.

Manning C., Schütze H. (1999) - *Foundations of Statistical Natural Language Processing,* MIT Press. Cambridge.

Menard N. (1983) - Mesure de la richesse lexicale, théorie et vérifications expérimentales, Slatkine-Champion, Paris.

Muller C. (1964) - Essai de statistique lexicale : L'illusion comique de P. Corneille, Klincksieck, Paris.

Muller C. (1968) - Initiation à la statistique linguistique, Larousse, Paris.

Muller C. (1977) - Principes et méthodes de statistique lexicale, Hachette, Paris.

Muller C.(1967) - Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille, Paris, Larousse.

Pêcheux M. (1969) - Analyse automatique du discours, Dunod, Paris.

Peschanski D. (1988) - Et pourtant, ils tournent. Vocabulaire et stratégie du PCF (1934 - 1936), Klincksieck, Paris.

Petruszewycz M. (1973) - L'histoire de la loi d'Estoup-Zipf, Math. Sciences Hum., n°44.

Pierrel J.-M.(2000) - *Ingénierie des langues*, Traité IC2 -Série informatique et SI, Hermes

Reinert M. (1990) - Alceste, Une méthodologie d'analyse des données textuelles et une Application : Aurélia de Gérard de Nerval, Bull. de Méthod. Sociol. n°26, p 24-54.

Romeu L. (1992) - Approche du discours éditorial de Ya et Arriba (1939 - 1945), Thèse Paris 3.

Salem A. (1984) - La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, Les Cahiers de l'Analyse des Données, Vol IX, n° 4, p 489-500.

Salem A. (1986) - Segments répétés et analyse statistique des données textuelles, Etude quantitative à propos du père Duchesne de Hébert, Histoire & Mesure, Vol. I- n° 2, Paris, Ed. du CNRS.

Salem A. (1987) - Pratique des segments répétés, Essai de satistique textuelle, Klincksieck, Paris.

Salem A. (1993) - Méthodes de la statistique textuelle, Thèse d'Etat, Université Sorbonne Nouvelle (Paris 3).

Sekhraoui M. (1981) - La saisie des textes et le traitement des mots: Problèmes posés, essai de solution, Mémoire, Ecole des hautes études en sciences sociales, Paris.

Tournier M. (1980) - D'ou viennent les fréquences de vocabulaire?, Mots N°1, p 189-212.

Tournier M. (1985a) - Sur quoi pouvons-nous compter ? Hommage à Hélène Nais, Verbum.

Tournier M. (1985b) - Texte propagandiste et cooccurrences. Hypothèses et méthodes pour l'étude de la sloganisation, Mots N°11, p 155-187.

Van Rijckevorsel J. (1987) - The application of fuzzy coding and horseshoes in multiple correspondances analysis, DSWO Press, Leyde.

Véronis J.(2000) - « Annotation automatique de corpus : panorama et état de la technique », *Ingénierie des langues*. J. M. Pierrel. Paris, Hermès.

Yule G.U. (1944) - The Statistical Study of Literary Vocabulary, Cambridge University Press, Reprinted in 1968 by Archon Books, Hamden, Connecticut.

Zipf G. K. (1935) - The Psychobiology of Language, an Introduction to Dynamic Philology, Boston, Houghton-Mifflin.

Cyber-bibliographie

Liens

- ?? FRANTEXT : http://zeus.inalf.cnrs.fr
- ?? LEXICOMETRICA : http://www.cavi.univ-paris3.fr/lexicometrica/
- ?? MARGES-LINGUISTIQUES : http://www.marges-linguistiques.com/
- ?? ATALA : http://www.atala.org/

Outils

- ?? HYPERBASE : http://lolita.unice.fr/pub/hyperbase/
- ?? TROPES : http://www.acetic.fr/
- ?? SPHINX : http://www.lesphinx-developpement.fr/
- ?? SPAD-T : http://www.cisia.com/
- ?? ALCESTE : http://www.image.cict.fr/
- ?? TALTAC : http://www.taltac.it/